



DEPARTMENT OF
COMPUTER SCIENCE

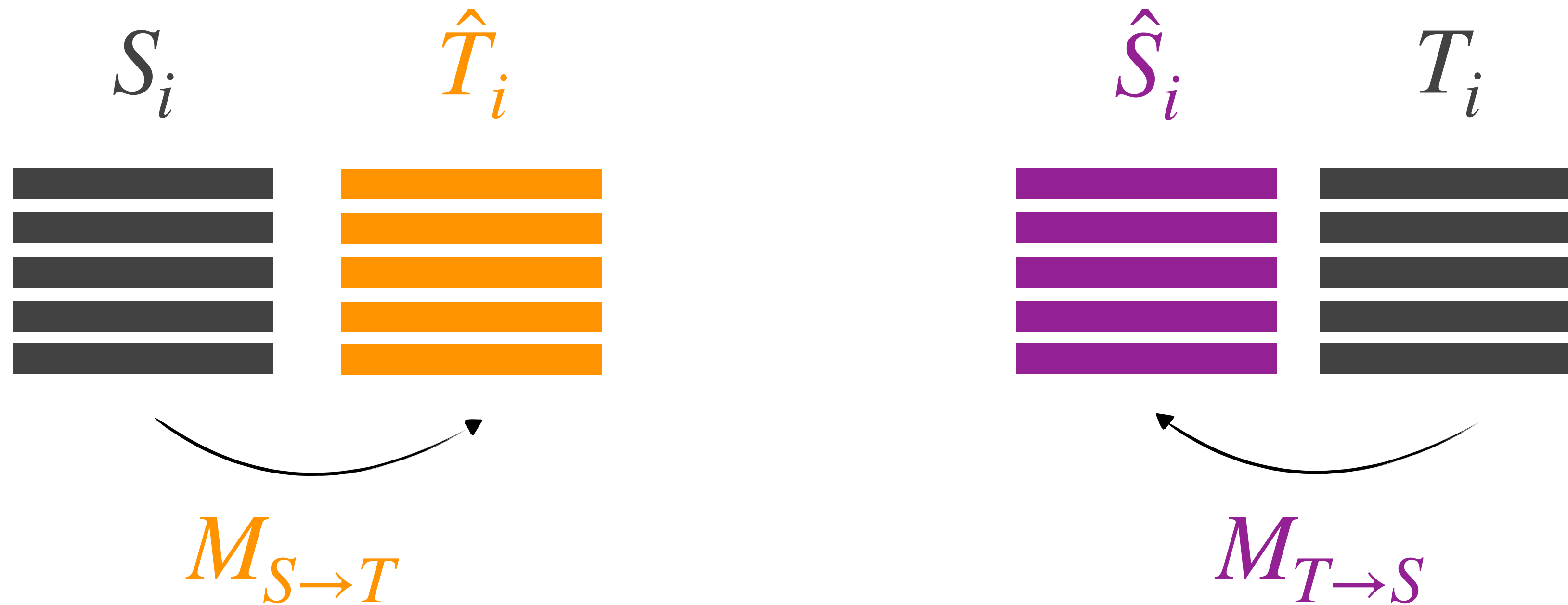
Can Synthetic Translations Improve Bitext Quality?

Eleftheria Briakou & Marine Carpuat

ebriakou@cs.umd.edu, marine@cs.umd.edu

Synthetic Translations

Translations generated by MT



Synthetic Translations

Improve Translation Quality of MT

Synthetic Translations

Improve Translation Quality of MT

AUGMENTING ORIGINAL TRANSLATIONS

- Forward translation
[Zhang and Zhong, 2016]
- Backward translation
[Sennrich et al., 2016]
- Data Diversification
[Nguyen et al., 2020]

Synthetic Translations

Improve Translation Quality of MT

AUGMENTING ORIGINAL TRANSLATIONS

- Forward translation
[Zhang and Zhong, 2016]
- Backward translation
[Sennrich et al., 2016]
- Data Diversification
[Nguyen et al., 2020]

REPLACING ORIGINAL TRANSLATIONS

- Sequence Distillation
[Gu et al., 2018]
- Data Rejuvenation
[Jiao et al., 2020]

Synthetic Translations

Improve Translation Quality of MT

Where does this improvement come from?

Synthetic Translations

Improve Translation Quality of MT

Where does this improvement come from?

Do synthetic translations...

- help language modeling?
- reinforce dominant patterns in the data?

Synthetic Translations

Improve Translation Quality of MT

Where does this improvement come from?

WE HYPOTHESIZE:

Synthetic translations are of higher quality than the original naturally occurring bitext

WE CONTRIBUTE:

An extensive empirical evaluation of the quality of bitext revised with synthetic translations

Synthetic Translations

Improve Translation Quality of MT

Where does this improvement come from?

WE HYPOTHESIZE:

Synthetic translations are of higher quality than the original naturally occurring bitext

WE CONTRIBUTE:

An extensive empirical evaluation of the quality of bitext revised with synthetic translations

Naturally Occurring Bitexts are treated as Exact Translations

GLOSS They had the largest population of dugongs in the area.

EL Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.

EN They had the largest population of dugons in the area.

no meaning differences

Naturally Occurring Bitexts are not always Exact Translations

fine-grained meaning differences

GLOSS Karavas was one of the settlements they created.

EL Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.

EN One of the first towns to be created was Vila Barreto.

no meaning differences

GLOSS They had the largest population of dugongs in the area.

EL Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.

EN They had the largest population of dugons in the area.

Naturally Occurring Bitexts are not always Exact Translations

fine-grained meaning differences

GLOSS Karavas was one of the settlements they created.

EL Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.

EN One of the first towns to be created was Vila Barreto.

no meaning differences

GLOSS They had the largest population of dugongs in the area.

EL Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.

EN They had the largest population of dugons in the area.

coarse meaning differences

GLOSS Hurricanes is a common phenomenon.

EL Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.

EN It is rare: There were only 10 known cases in 1998.

Can Synthetic Translations Improve Bitext Quality?

GLOSS Karavas was one of the settlements they created.

EL Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.

EN One of the first towns to be created was Vila Barreto.

GLOSS They had the largest population of dugongs in the area.

EL Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.

EN They had the largest population of dugons in the area.


GLOSS Hurricanes is a common phenomenon.

EL Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.

EN It is rare: There were only 10 known cases in 1998.

Synthetic Translations

Can Revise Meaning Mismatches in Bitexts

GLOSS Karavas was one of the settlements they created.		ST	One of settlements to be created was Karavas.
EL Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.		EN	One of the first towns to be created was Vila Barreto.
GLOSS They had the largest population of dugongs in the area.			
EL Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.		EN	They had the largest population of dugons in the area.
GLOSS Hurricanes is a common phenomenon.			
EL Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.		EN	It is rare: There were only 10 known cases in 1998.

Synthetic Translations can Introduce Meaning Mismatches in Bitext

GLOSS Karavas was one of the settlements they created.

Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.

One of settlements to be created was Karavas.

One of the first towns to be created was Vila Barreto.

GLOSS They had the largest population of dugongs in the area.

EL Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.



ST

They had the largest population of alikers in the area.

EN

They had the largest population of dugons in the area.

GLOSS Hurricanes is a common phenomenon.

EL Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.

EN

It is rare: There were only 10 known cases in 1998.

Synthetic Translations can Improve Bitext Quality

Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.

One of settlements to be created was Karavas.

One of the first towns to be created was Vila Barreto.

They had the largest population of alikers in the area.

Είχαν το μεγαλύτερο πληθυσμό αλικόρων στην περιοχή.

They had the largest population of dugons in the area.

Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.

The appearance of hurricanes is a common phenomenon.

It is rare: There were only 10 known cases in 1998.

Synthetic Translations can Improve Bitext Quality

UNDER SEMANTIC CONTROLS



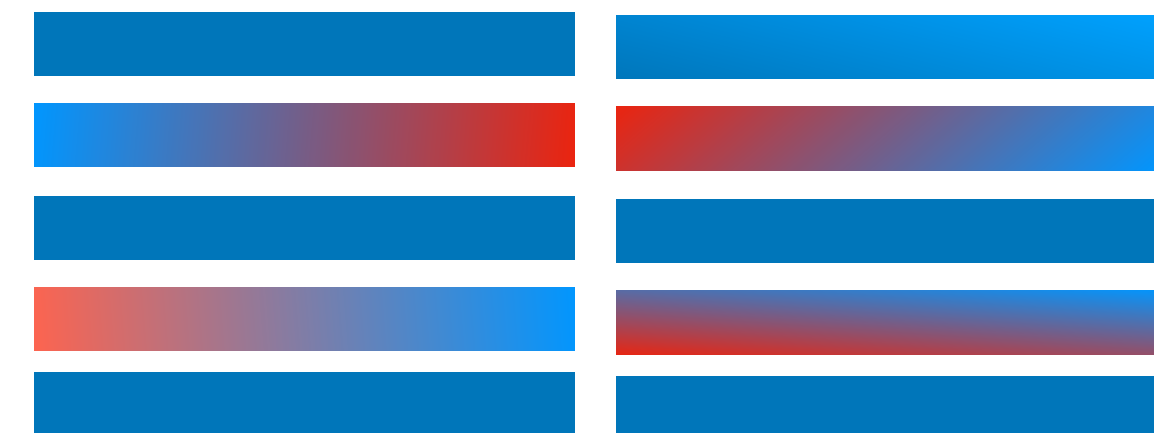
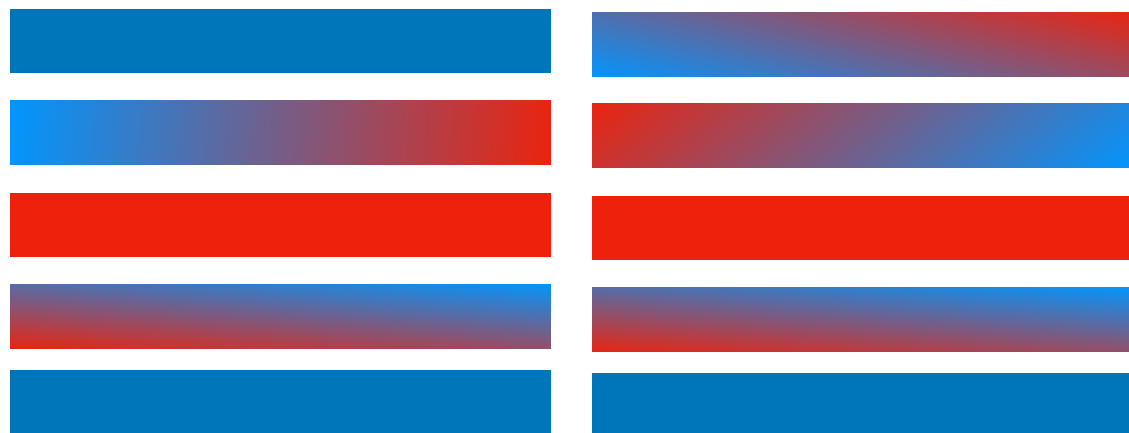
Revising Bitext with Selective Replacement of Synthetic Translations

Original Corpus

Revised Corpus

(S, T)

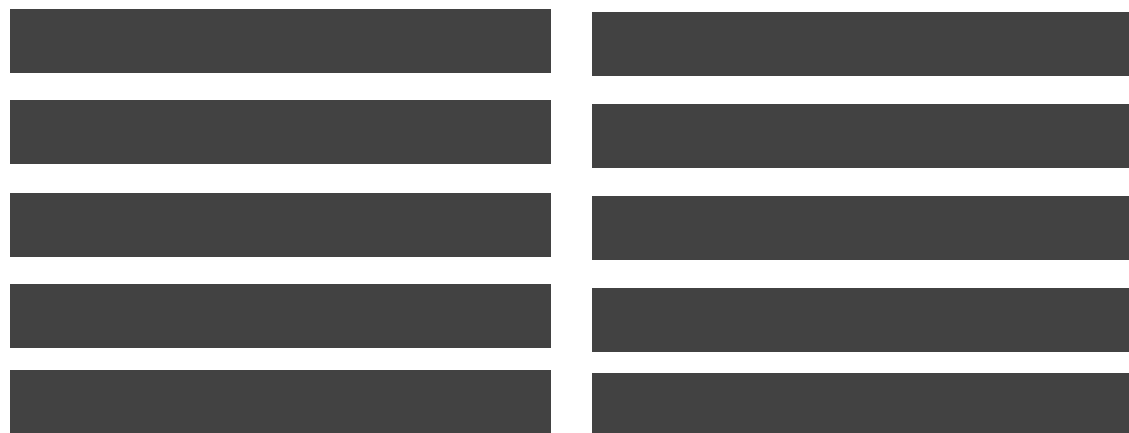
(\hat{S}, \hat{T})



Revising Bitext with Selective Replacement of Synthetic Translations

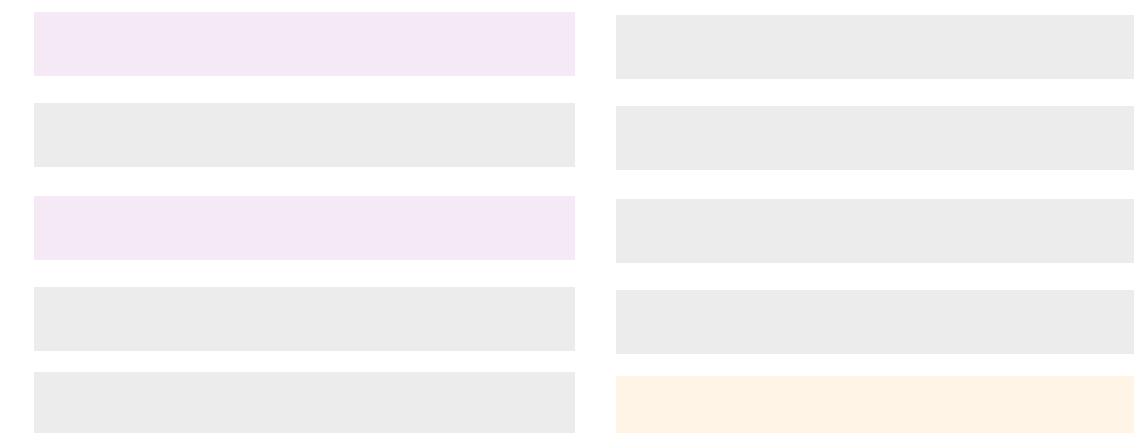
Original Corpus

(S, T)

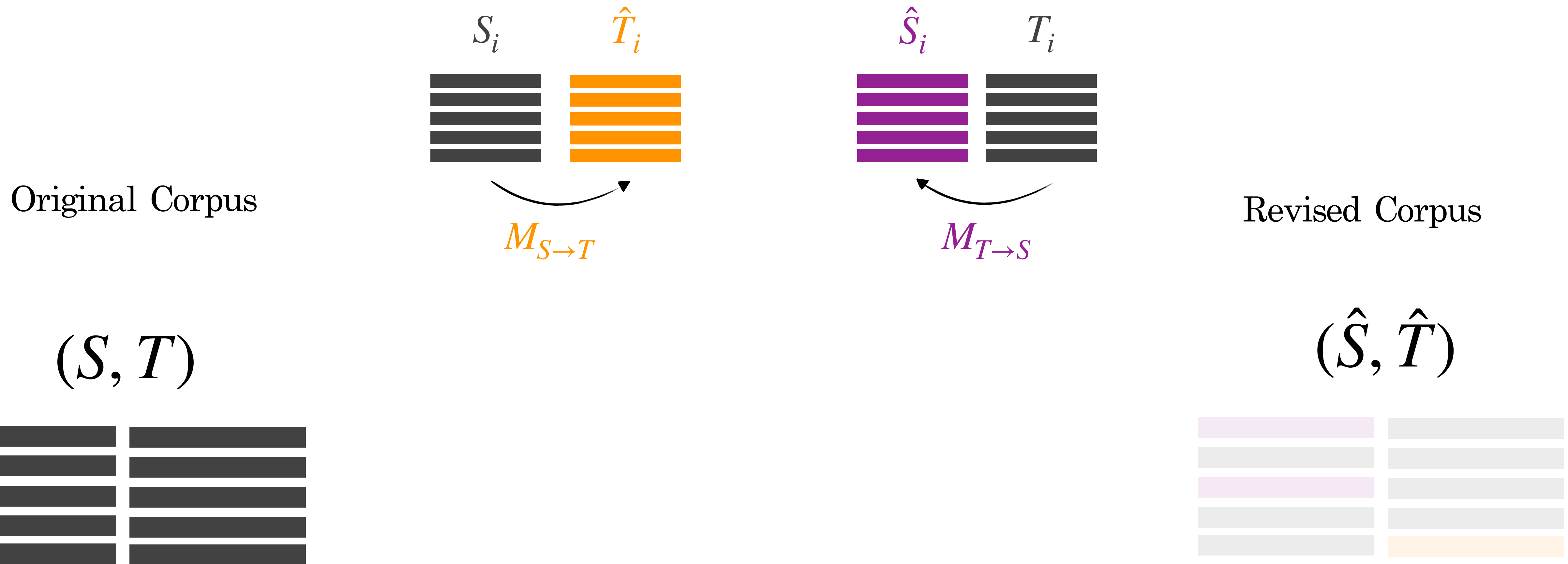


Revised Corpus

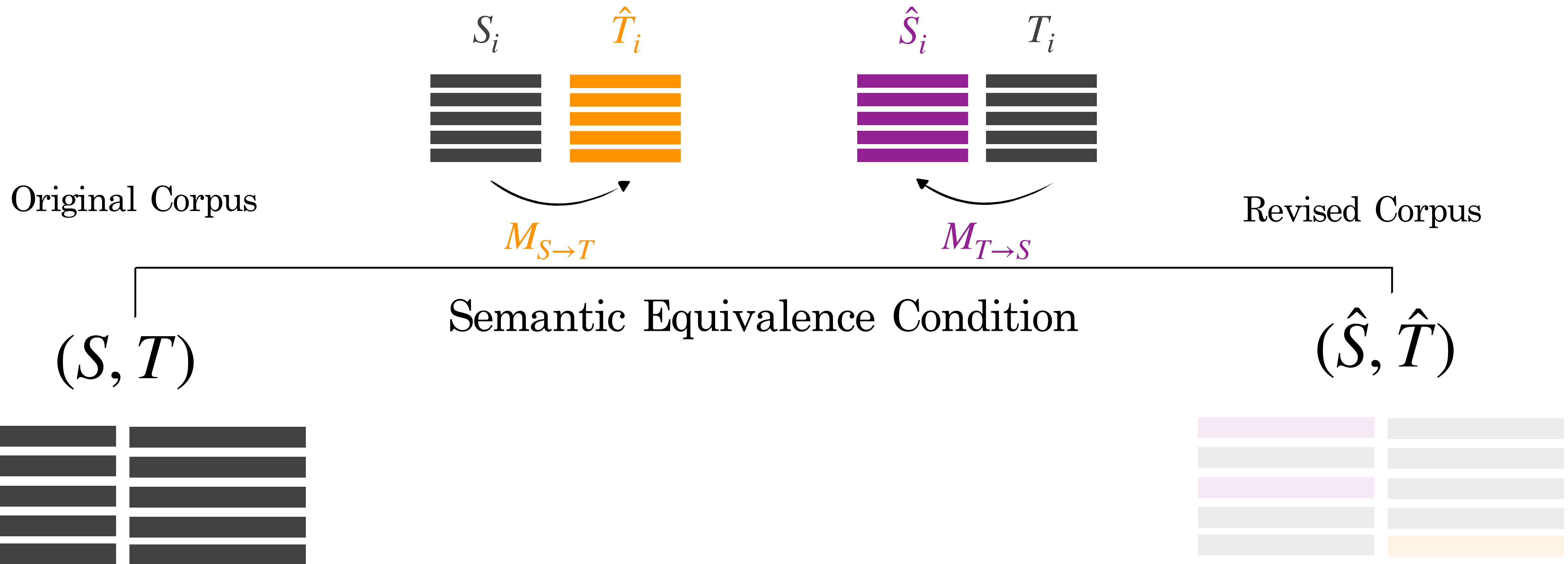
(\hat{S}, \hat{T})



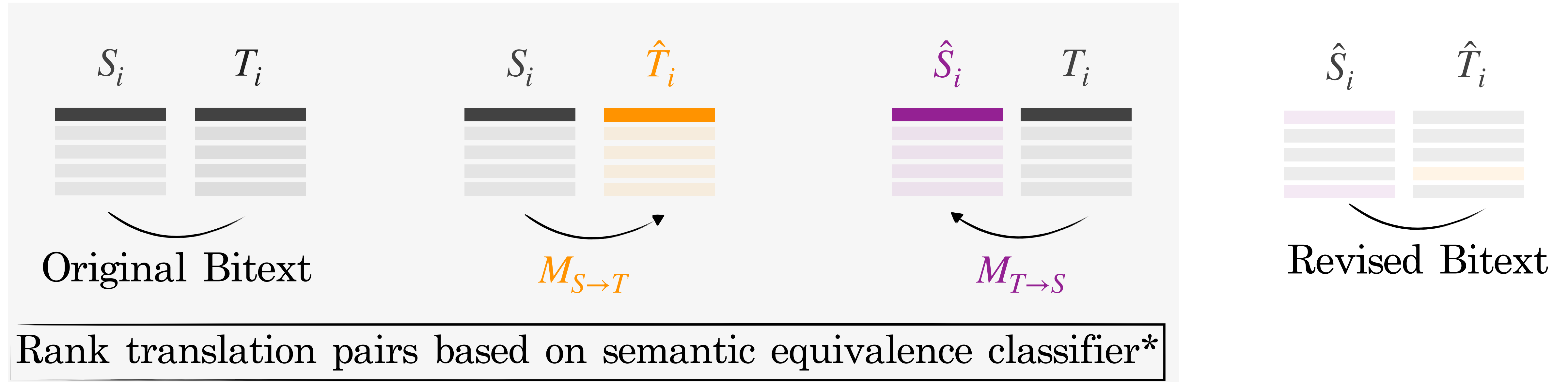
Revising Bitext with Selective Replacement of Synthetic Translations



Revising Bitext with Selective Replacement of Synthetic Translations

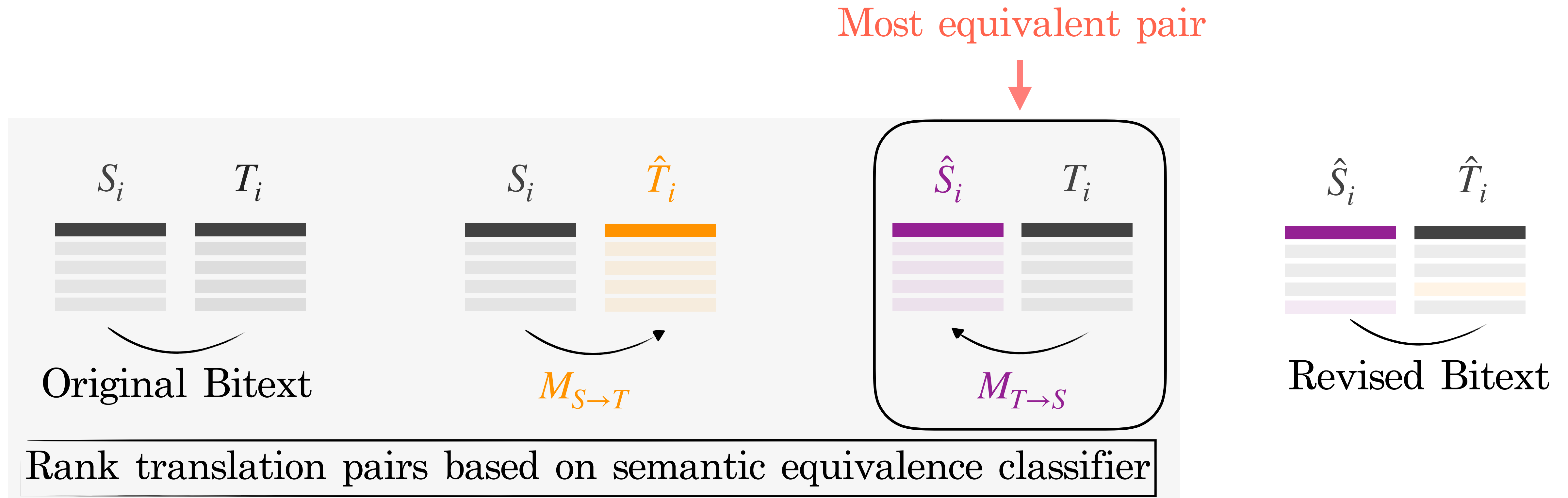


Revising Bitext with Selective Replacement of Synthetic Translations



*Eleftheria Briakou & Marine Carpuat. 2020. [Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank](#). In EMNLP

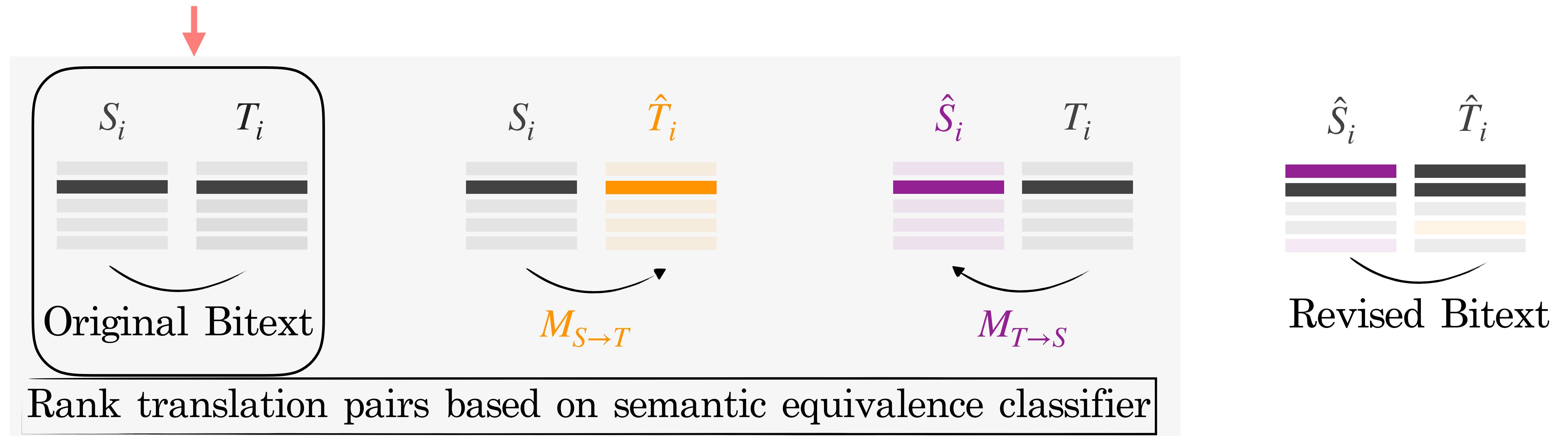
Synthetic translations replace the original **only if** they yield a more equivalent translation



*Eleftheria Briakou & Marine Carpuat. 2020. [Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank](#). In EMNLP

Synthetic translations replace the original **only if** they yield a more equivalent translation

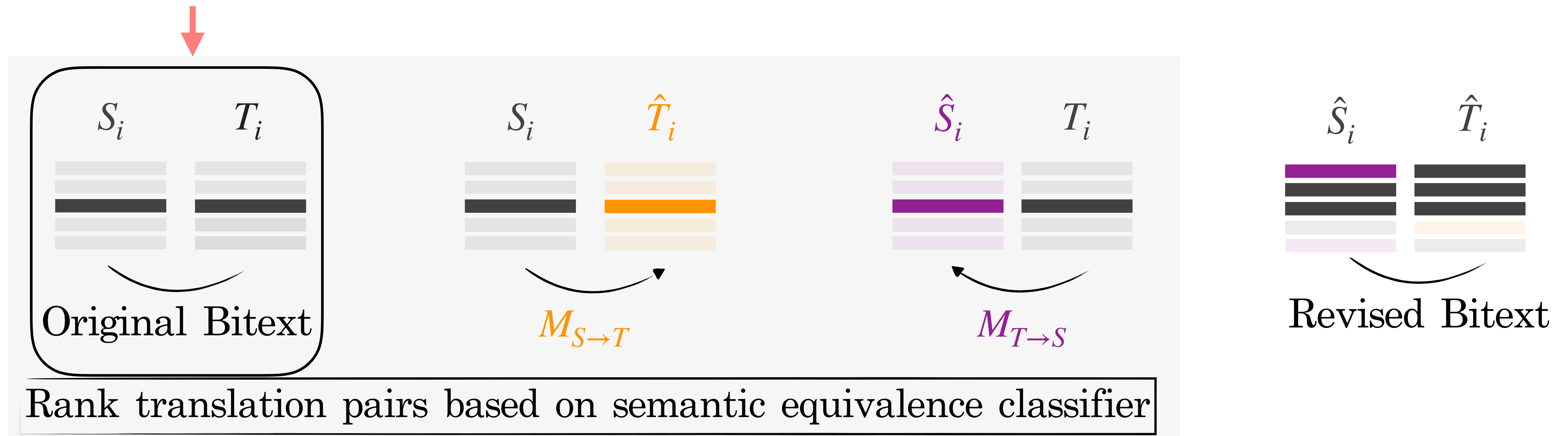
Most equivalent pair



*Eleftheria Briakou & Marine Carpuat. 2020. [Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank](#). In EMNLP

Synthetic translations replace the original **only if** they yield a more equivalent translation

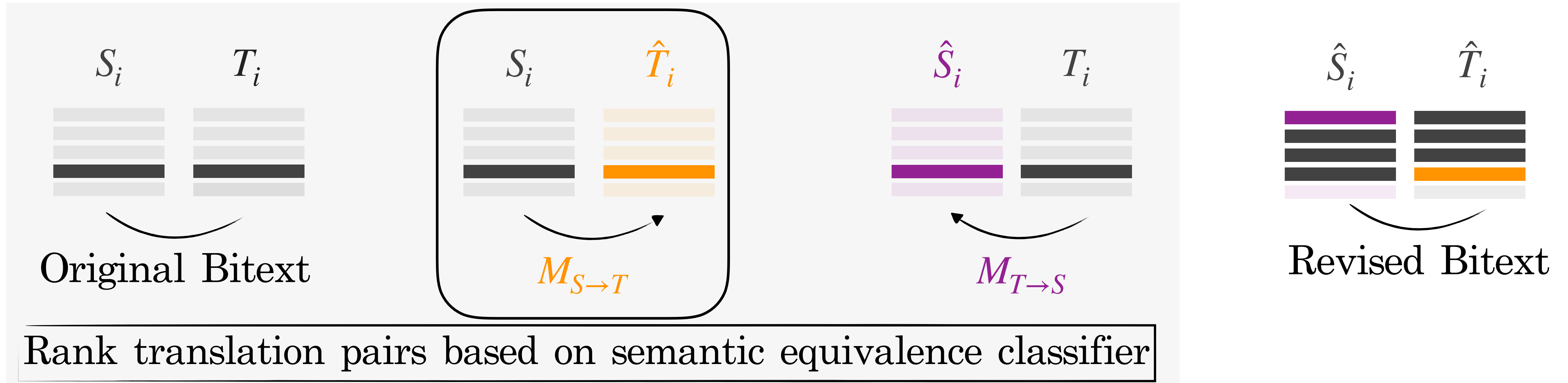
Most equivalent pair



*Eleftheria Briakou & Marine Carpuat. 2020. [Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank](#). In EMNLP

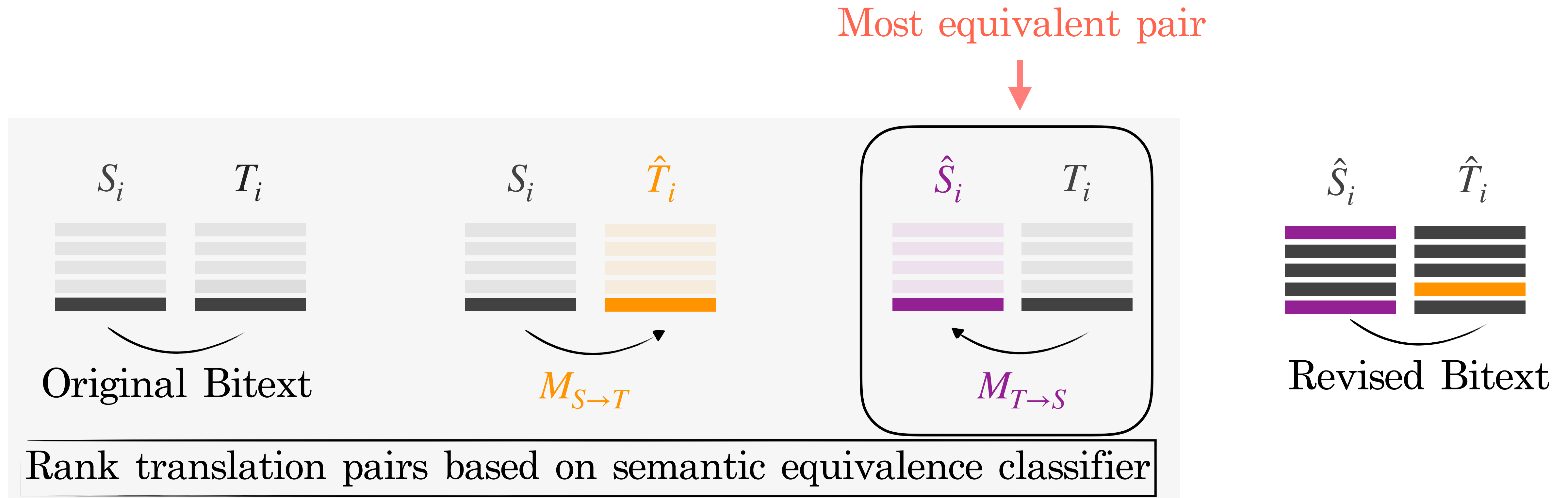
Synthetic translations replace the original **only if** they yield a more equivalent translation

Most equivalent pair



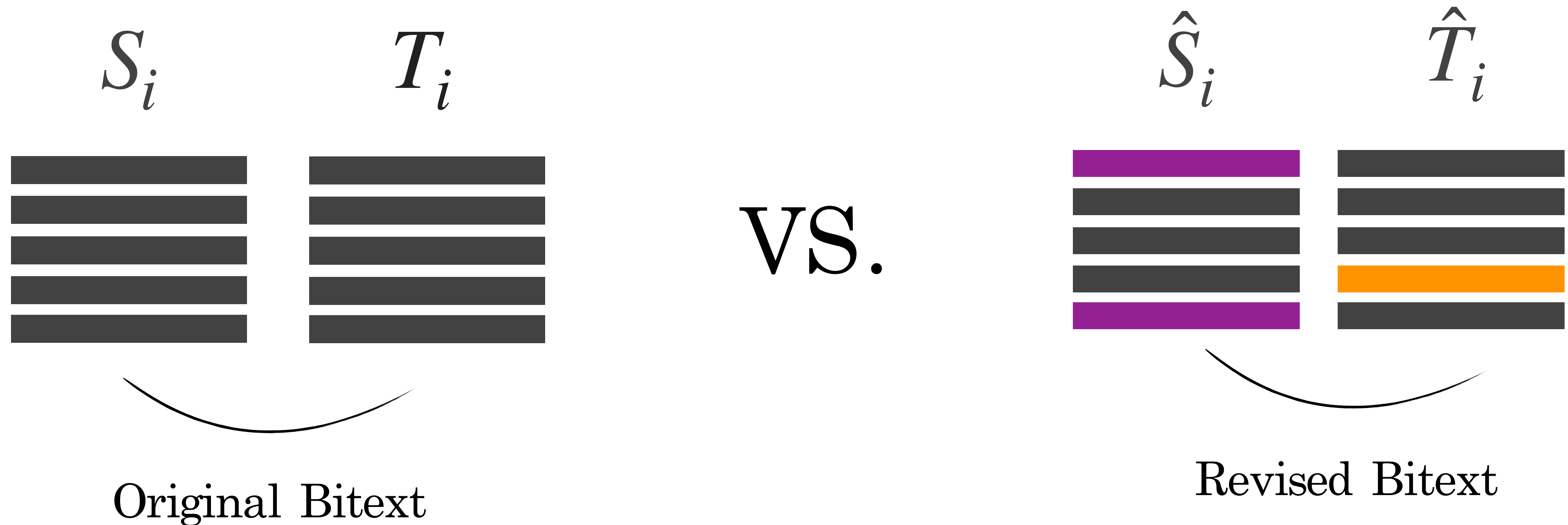
*Eleftheria Briakou & Marine Carpuat. 2020. [Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank](#). In EMNLP

Synthetic translations replace the original **only if** they yield a more equivalent translation



*Eleftheria Briakou & Marine Carpuat. 2020. [Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank](#). In EMNLP

Can Synthetic Translations Improve Bitext Quality?



Original vs. Revised Bitext: An Empirical Evaluation of Bitext Quality

Intrinsic:

Extrinsic:

Original vs. Revised Bitext: An Empirical Evaluation of Bitext Quality

Intrinsic: Human Assessments of Equivalence

Extrinsic:

Original vs. Revised Bitext:

An Empirical Evaluation of Bitext Quality

Intrinsic: Human Assessments of Equivalence

Extrinsic: Performance on downstream NLP tasks

Original vs. Revised Bitext: An Empirical Evaluation of Bitext Quality

Intrinsic: Human Assessments of Equivalence

Extrinsic: Performance on downstream NLP tasks

➔ Bilingual Lexicon Induction via word alignment

↑ Bitext Quality ➔ more accurate cross-lingual lexical mappings

Original vs. Revised Bitext: An Empirical Evaluation of Bitext Quality

Intrinsic: Human Assessments of Equivalence

Extrinsic: Performance on downstream NLP tasks

→ Bilingual Lexicon Induction via word alignment

↑ Bitext Quality → more accurate cross-lingual lexical mappings

→ Machine Translation [WMT Parallel Corpus Filtering evaluation]

↑ Bitext Quality → more reliable training signal

Bitext Quality Evaluations: Experimental Settings

- ✓ Training bitexts → WikiMatrix (mined)
- ✓ Language-pairs → Greek-English (EL-EN) [~750K]
Romanian-English (RO-EN) [~600K]
- ✓ MT Test Sets → TED
- ✓ BLI Test Sets → MUSE

Medium Resource Focus:

(a) Sufficient MT Quality

(b) Bitext Improvement needed

Intrinsic Evaluation: Human Assessments of Equivalence



- ▶ 100 samples
- ▶ 3 annotators
- ▶ Lang: EL-EN

Intrinsic Evaluation: Human Assessments of Equivalence

Source (original)

Ένας από τους οικισμούς που δημιουργήσαν ήταν ο Καραβάς.



- ▶ 100 samples
- ▶ 3 annotators
- ▶ Lang: EL-EN

Intrinsic Evaluation: Human Assessments of Equivalence

Source (original)

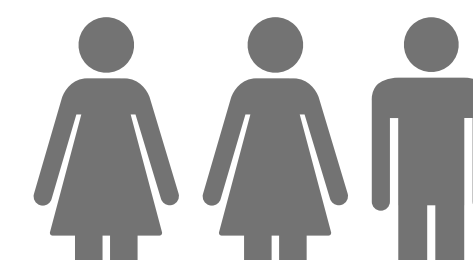
Ένας από τους οικισμούς που δημιουργήσαν ήταν ο Καραβάς.

Target A (original)

One of the first towns to be created was Vila Barreto.

Target B (revised)

One of settlements to be created was Karavas.



- ▶ 100 samples
- ▶ 3 annotators
- ▶ Lang: EL-EN

Intrinsic Evaluation: Human Assessments of Equivalence

“Which sentence conveys the meaning of the source better?”

Source (original)

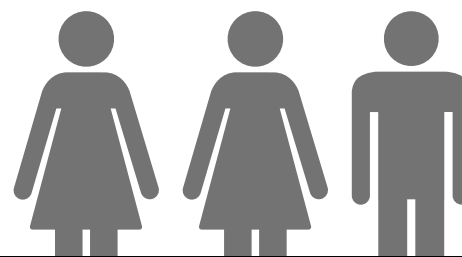
Ένας από τους οικισμούς που δημιουργήσαν ήταν ο Καραβάς.

Target A (original)

One of the first towns to be created was Vila Barreto.

Target B (revised)

One of settlements to be created was Karavas.



- ▶ 100 samples
- ▶ 3 annotators
- ▶ Lang: EL-EN

Intrinsic Evaluation: Synthetic Translations improve Bitext Quality

“Which sentence conveys the meaning of the source better?”

Source (original)

Ένας από τους οικισμούς που δημιουργήσαν ήταν ο Καραβάς.

Target A (original)

One of the first towns to be created was Vila Barreto.

→ 12%

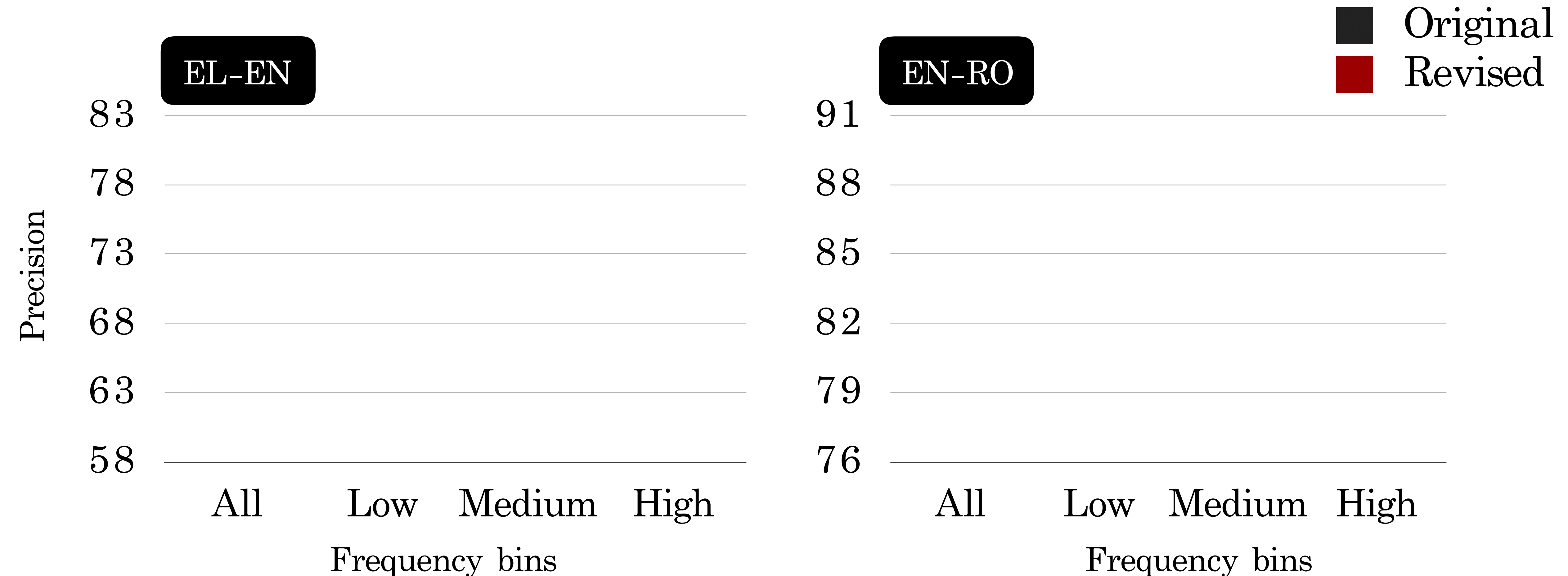
Target B (revised)

One of settlements to be created was Karavas.

→ 88%

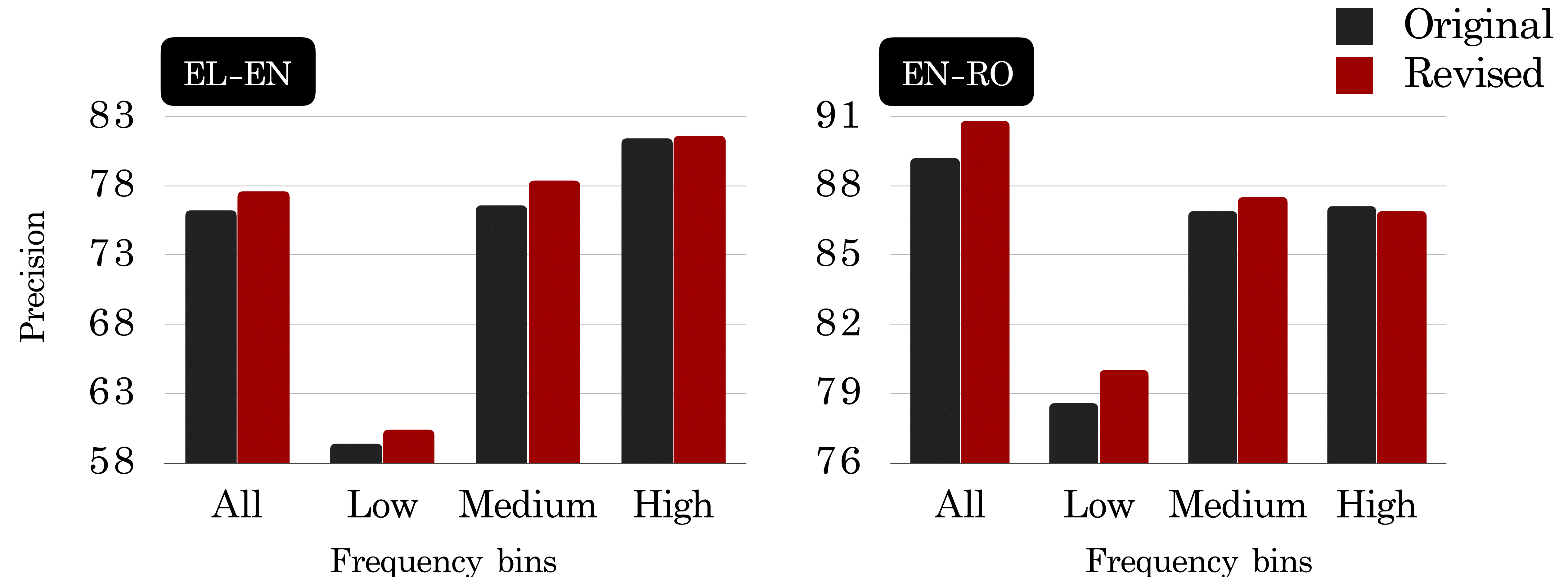
Extrinsic Evaluation [BLI]

Do revised bitexts induce more precise lexicons?



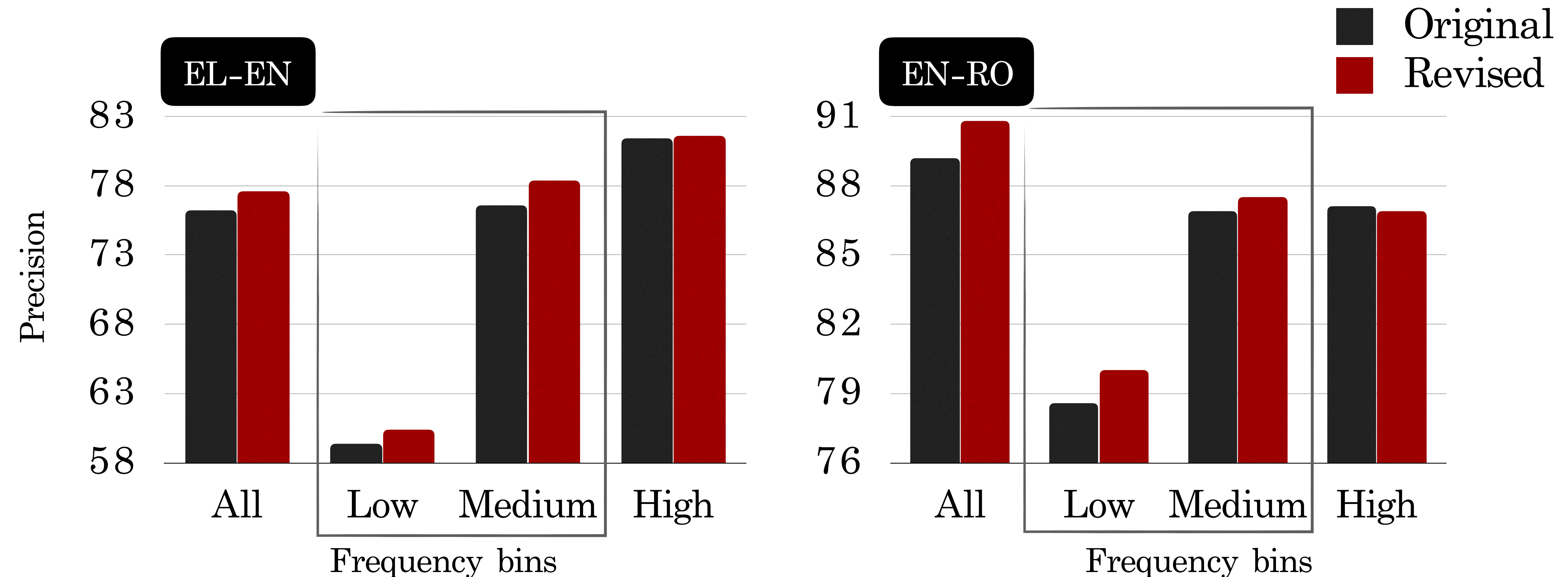
Extrinsic Evaluation [BLI]

Revised bitext yields more precise lexicons for low-medium frequency words



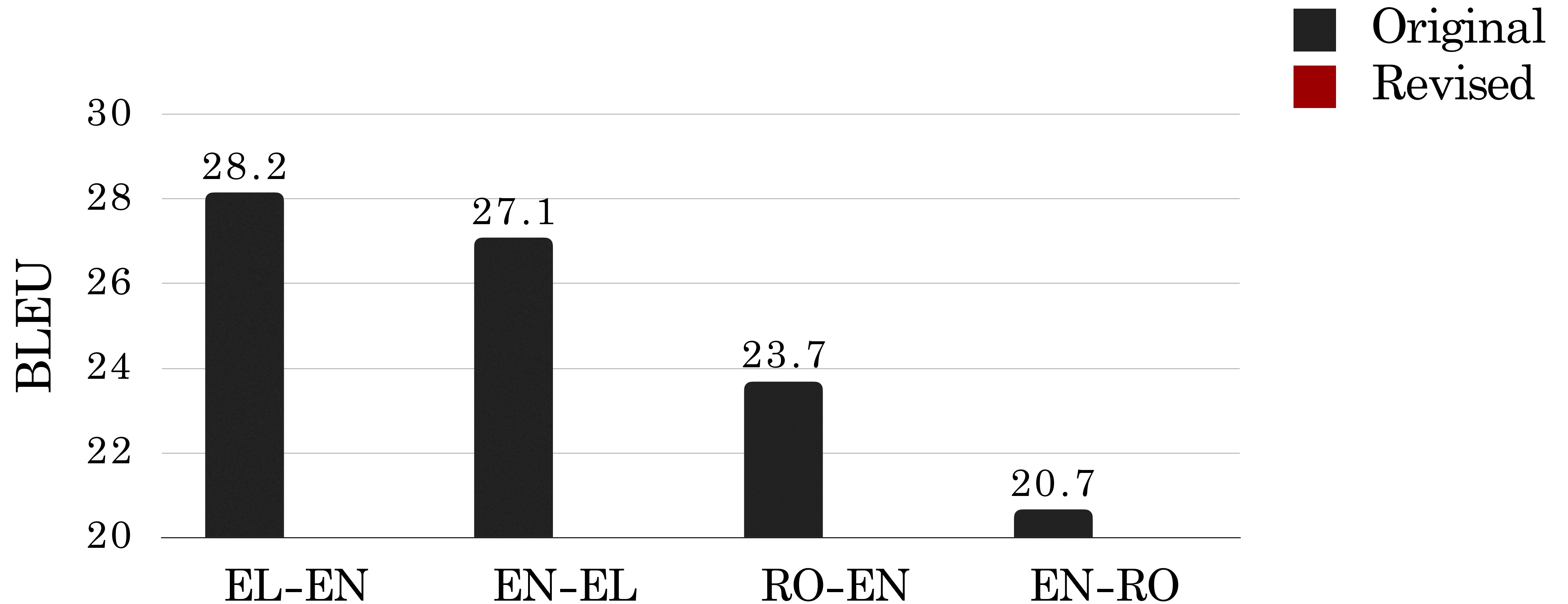
Extrinsic Evaluation [BLI]

Revised bitext yields more precise lexicons for low-medium frequency words



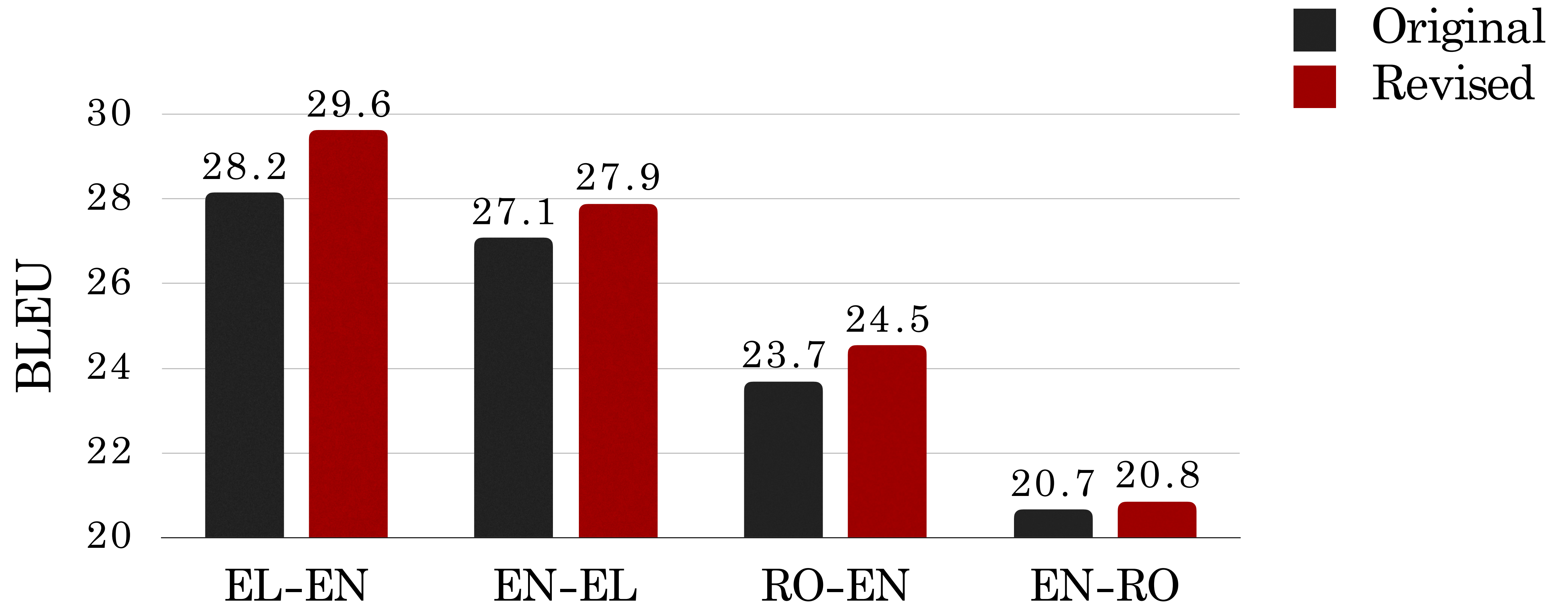
Extrinsic Evaluation [MT from scratch]

Do revised bitexts improve MT quality?



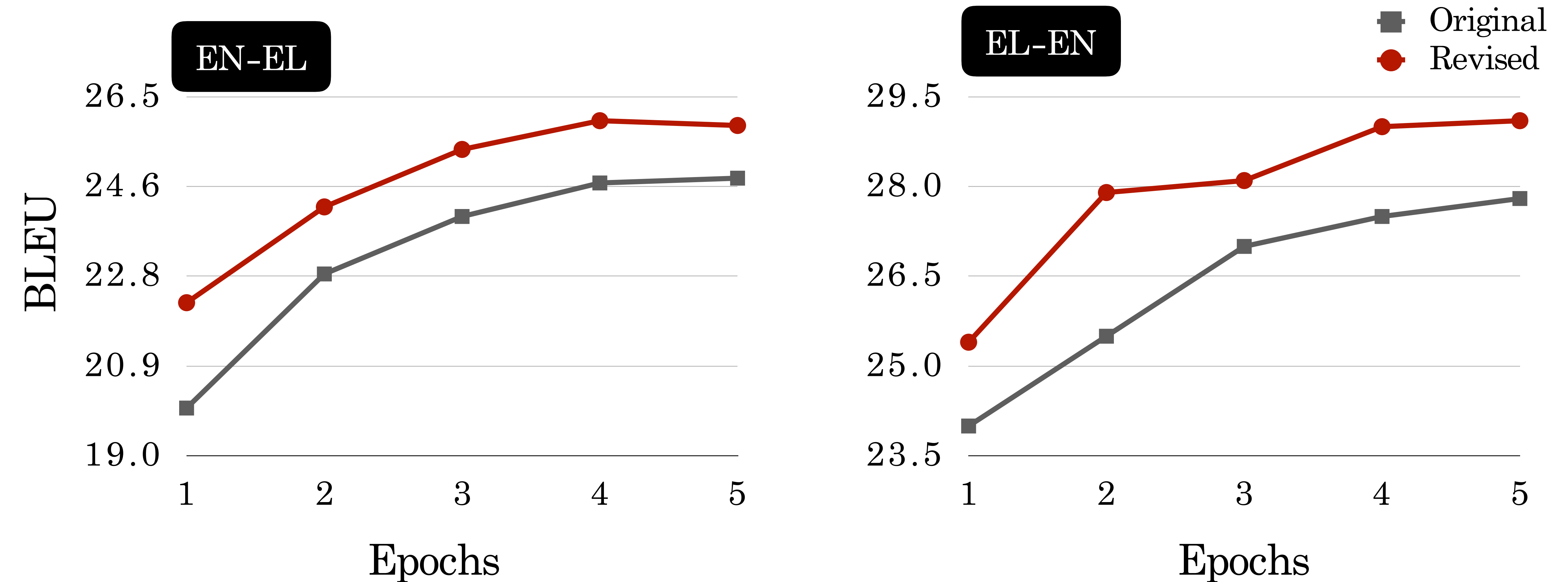
Extrinsic Evaluation [MT from scratch]

Revised bitext yields better translation quality than training on the original



Extrinsic Evaluation [MT continued training]

Revised bitext yields better translation quality than training on the original



Can Synthetic Translations Improve Bitext Quality?

Can Synthetic Translations Improve Bitext Quality?

Yes, when...

they selectively replacing imperfect translations in naturally occurring bitexts under a semantic equivalence condition

Can Synthetic Translations Improve Bitext Quality?

Yes, when...

they selectively replacing imperfect translations in naturally occurring bitexts under a semantic equivalence condition

According to...

intrinsic evaluations of semantic equivalence and extrinsic evaluations on BLI and MT tasks

Can Synthetic Translations Improve Bitext Quality?

Yes, when...

they selectively replacing imperfect translations in naturally occurring bitexts under a semantic equivalence condition

According to...

intrinsic evaluations of semantic equivalence and extrinsic evaluations on BLI and MT tasks